Preprocessing for Affymetrix GeneChip Data

1/18/2011
Copyright © 2011 Dan Nettleton

Affymetrix .CEL Files

- A .CEL file contains one number representing signal intensity for each probe cell on a single GeneChip.
- .CEL files can be read with Affymetrix software or in R using the Bioconductor package affy.
- We will discuss two methods for normalizing and obtaining expression measures using data from Affymetrix .CEL files.

Methods

2. Robust Multi-array Average (RMA) proposed by Irizarry et al. (2003) Biostatistics 4, 249-264.

These are perhaps the two most popular of many methods for normalizing and computing expression measures using Affymetrix data. Currently > 50 methods are described and compared at http://affycomp.biostat.jhsph.edu/.

MAS 5.0 Signal: Background Adjustment

- Each chip is divided into 16 rectangular zones.
- The lowest 2% of intensities in each zone are averaged to form a zone-specific background value denoted bZk for zones k=1, 2, ..., 16.
- The standard deviation of the lowest 2% of intensities in each zone is calculated and denoted nZk for zones k=1, 2, ..., 16.
- Let d_k(x,y) denote the distance from the center of zone k to a probe cell located at coordinates (x,y) on the chip.
MAS 5.0 Signal: Background Adjustment (continued)

- Let \( w_k(x,y) = \frac{1}{d_k(x,y) + 100} \).
- Denote the background for the cell located at coordinates \((x,y)\) by \( b(x,y) = \frac{\sum_{k=1}^{n} w_k(x,y) b \zeta_k}{\sum_{k=1}^{n} w_k(x,y)} \).
- Denote the "noise" for the cell located at coordinates \((x,y)\) by \( n(x,y) = \frac{\sum_{k=1}^{n} w_k(x,y) n \zeta_k}{\sum_{k=1}^{n} w_k(x,y)} \).

- Let \( I(x,y) \) denote the original intensity of the cell located at coordinates \((x,y)\) on the chip. (75th percentile of 36 pixel intensities in the center of the cell.)
- Let \( I'(x,y) = \max (I(x,y), 0.5) \).
- Define the background-adjusted intensity for the cell at coordinates \((x,y)\) by \( A(x,y) = \max \{ I'(x,y) - b(x,y), 0.5n(x,y) \} \).
- Henceforth these background-adjusted intensities will be referred to as either PM or MM for perfect match or mismatch cells, respectively.

MAS 5.0 Signal: Ideal Mismatch Computation

- MM values are supposed to provide measures of cross-hybridization and stray signal intensity that inflate the value of PM.
- In the simplest case, a PM value would be corrected simply by subtracting its corresponding MM value.
- However, some MM values are bigger than their corresponding PM values so that PM-MM would become negative.
- Because negative values do not make a lot of sense and would pose problems with subsequent steps in analysis, Affymetrix determines an Ideal Mismatch (IM) value for each probe pair that is guaranteed to be less than PM.

MAS 5.0 Signal: Ideal Mismatch Computation (continued)

- If \( PM_j > MM_j \), then \( IM_j = MM_j \) and no further computation is needed.
- If \( PM_j \leq MM_j \), compute \( M = \frac{TBW \{ \log_2(PM_1/MM_1), \ldots, \log_2(PM_n/MM_n) \} }{1 + (0.03 - M)^{10/0.03}} \)
  where \( TBW \) denotes a one-step Tukey BiWeight (a special weighted average described later).

MAS 5.0 Signal: Signal Log Value Computation

- Let \( V_j = \max (PM_j - IM_j, 2^{-20}) \).
- Define the probe value for the \( j^{th} \) probe pair by \( PV_j = \log_2(V_j) \).
- The signal log value for a given probe set is defined by \( SLV = TBW (PV_1, PV_2, \ldots, PV_n) \)
  where \( TBW \) denotes a one-step Tukey BiWeight (a special weighted average to be discussed later).
• Let SLVi denote the signal log value for the ith probe set on a single chip.
• Let I denote the number of probe sets on the chip.
• Let SF = 500/TrimMean( 2SLVi, 2SLVi, ..., 2SLVi ; 0.02, 0.98).
  The average of the values in parentheses
  that are strictly between the 0.02 and 0.98
  quantiles of the values in parentheses.
• MAS 5.0 Signal for the ith probe set is Signali = SF * 2SLVi.
• All computations are done separately for each chip to
  obtain a Signal value for each chip and probe set.

The One-Step Tukey BiWeight Estimator
Used by Affymetrix

Recall the bisquare weight function defined as

\[ B(t) = \begin{cases} 
(1 - t^2)^2 & \text{for } |t| < 1 \\
0 & \text{for } |t| \geq 1. 
\end{cases} \]

TBW( x1, x2, ..., xn ) = \sum_{i=1}^{n} B(t_i) x_i \\
\sum_{i=1}^{n} B(t_i)

An Example

Compute TBW ( 1, 7, 13, 15, 28, 1075 ).

\[ m = \left( \frac{13 + 15}{2} \right) = 14. \]

\[ \text{MAD} = \text{median} ( |1-14|, |7-14|, |13-14|, |15-14|, |28-14|, |1075-14| ) = \left( \frac{7 + 13}{2} \right) = 10. \]

\[ t_1 = \frac{-13}{50}, \quad t_2 = \frac{-7}{50}, \quad t_3 = \frac{-1}{50}, \quad t_4 = \frac{1}{50}, \quad t_5 = \frac{14}{50}, \quad t_6 = \frac{1061}{50} \]

\begin{align*}
B(t_1) &= B(0.26) = (1 - 0.26^2)^2 = 0.8693698 \\
B(t_2) &= B(0.14) = (1 - 0.14^2)^2 = 0.9611842 \\
B(t_3) &= B(0.02) = (1 - 0.02^2)^2 = 0.9992002 \\
B(t_4) &= B(0.02) = (1 - 0.02^2)^2 = 0.9992002 \\
B(t_5) &= B(0.28) = (1 - 0.28^2)^2 = 0.8493466 \\
B(t_6) &= 0
\end{align*}

0.8693698*1 + 0.9611842*7 + 0.9992002*13 + 0.9992002*15 + 0.8493466*28 + 0*1075 = 12.68772.

Obtaining MAS5.0 Signal Values from Affymetrix .CEL Files

• MAS5.0 Signal values can be obtained from Affymetrix software.

• Approximate MAS5.0 Signal values can be computed with the mas5 function that is part of the Bioconductor package affy.

• Use whichever method is easiest for you. The differences do not seem to be large enough to matter.
Bioconductor

- Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data.
- Bioconductor provides R packages useful for the analysis of gene expression data.
- Information about Bioconductor, including installation instructions, can be found at www.bioconductor.org.

R Commands for Obtaining MAS5.0 Signal Values from Affymetrix .CEL Files

```r
# Load the Bioconductor package affy.
library(affy)

# Set the working directory to the directory containing all the .CEL files.
setwd("C:/z/Courses/Smicroarray/AffyCel")

# Read the .CEL file data.
Data=ReadAffy()

# Compute the MAS5.0 Signal Values
signal=mas5(Data)

# Write the data to a tab-delimited text file.
write.exprs(signal, file="mydata.txt")
```

Robust Multi-array Average (RMA)

1. Background adjust PM values from .CEL files.
2. Take the base-2 log of each background-adjusted PM intensity.
3. Quantile normalize values from step 2 across all GeneChips.
4. Perform median polish separately for each probe set with rows indexed by GeneChip and columns indexed by probe.
5. For each row, find the average of the fitted values from step 4 to use as probe-set-specific expression measures for each GeneChip.

RMA: Background Adjustment

Assume \( PM = S + B \) where

- \( S \sim \text{Exp}(\lambda) \) independent of
- \( B \sim N^+(\mu, \sigma^2) \).

\( N^+(\mu, \sigma^2) \) denotes \( N(\mu, \sigma^2) \) truncated on the left at 0.

The Probability Density Function of the Exponential Distribution with Mean \( 1/\lambda = 10000 \)

\[ \lambda e^{-\lambda s} \]

\( s \)

The Probability Density Function of the Normal Distribution with Mean \( \mu = 1000 \) and Variance \( \sigma^2 = 300^2 \)

\[ \frac{e^{-\frac{(b-\mu)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{0.5}} \]

\( b \)
The Probability Density Function of $s + b$ where $s \sim \text{Exp}(\lambda = 1/10000)$ and $b \sim \mathcal{N}(\mu = 1000, \sigma^2 = 300^2)$

**RMA: Background Adjustment (continued)**

The Probability Density Function of $s + b$

\[
\text{Density of } s + b = \frac{1}{\lambda} e^{-\frac{s}{\lambda}} - \frac{1}{\sigma^2} (s - \mu)^2 e^{-\frac{1}{2} \left(\frac{s - \mu}{\sigma}\right)^2}.
\]

Separately for each chip, estimate $\mu$, $\sigma$, and $\lambda$ from the observed PM distribution. Plug those estimates into the formula above to obtain an estimate of $E(S|PM)$ for each PM value. These serve as background-adjusted PM values.

**Obtaining Estimates of $\mu$, $\sigma$, and $\lambda$** (unpublished description of the procedure)

- Estimate the mode of the PM distribution using a kernel density estimate of the PM density.
- Estimate the density of the PM values less than the mode. The mode of this distribution serves as an estimate of $\mu$.
- Assume the data to the left of the estimate of $\mu$ are the background observations that fell below their mean. Use those observations to estimate $\sigma$.
- Subtract the estimate of $\mu$ from all observations larger than the estimate. The mode of this distribution estimates $1/\lambda$.

**Density Estimate of PM Data below the Estimated Mode of the PM Distribution**

This data is used to estimate $\sigma$ as 642.3.

**Estimate of $\sigma$**

According to the RMA R code, $\sigma$ is estimated as follows:

\[
\hat{\sigma} = \sqrt{\frac{\sum_{PM < \hat{\mu}} (PM - \hat{\mu})^2}{\#\{PM < \hat{\mu}\} - 1}}
\]

The purpose of the factor of 2 in the numerator is not clear.
Density Estimate of PM – \( \hat{\mu} \) Values Greater than Zero

The mean of these values would be a much better estimate of \( 1/\lambda \) in this case.
(Mean is 9848 and \( 1/\lambda = 10000 \).)

Estimate of \( 1/\lambda = 2019 \)

RMA: Quantile Normalization
1. After background adjustment, find the smallest log2(PM) on each chip.
2. Average the values from step 1.
3. Replace each value in step 1 with the average computed in step 2.
4. Repeat steps 1 through 3 for the second smallest values, third smallest values, ..., largest values.

RMA: Median Polish
• For a given probe set with J probe pairs, let \( y_{ij} \) denote the background-adjusted, base-2-logged, and quantile-normalized value for GeneChip \( i \) and probe \( j \).
• Assume \( y_{ij} = \mu_i + \alpha_j + e_{ij} \) where \( \alpha_1 + \alpha_2 + ... + \alpha_n = 0 \).
  - gene expression of the probe set on GeneChip \( i \)
  - probe affinity affect for the \( j \) probe in the probe set
  - residual for the \( j \) probe on the \( i \)th GeneChip
• Perform Tukey’s Median Polish on the matrix of \( y_{ij} \) values with \( y_{ij} \) in the \( i \)th row and \( j \)th column.

RMA: Median Polish (continued)
• Let \( \hat{y}_{ij} \) denote the fitted value for \( y_{ij} \) that results from the median polish procedure.
• Let \( \hat{\alpha}_j = \hat{y}_{..} - \hat{y}_{..} \) where \( \hat{y}_{..} = \frac{\Sigma_{i=1}^{I} \hat{y}_{ij}}{I} \) and \( \hat{y}_{..} = \frac{\Sigma_{i=1}^{I} \Sigma_{j=1}^{J} y_{ij}}{IJ} \) and \( I \) denotes the number of GeneChips.
• Let \( \hat{\mu}_i = \hat{y}_{i.} = \Sigma_{j=1}^{J} \hat{y}_{ij} / J \)
  - \( \hat{\mu}_i \) is the probe-set-specific measure of expression for GeneChip \( i \).

An Example
Suppose the following are background-adjusted, log2-transformed, quantile-normalized PM intensities for a single probe set. Determine the final RMA expression measures for this probe set.

<table>
<thead>
<tr>
<th>Probe</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneChip 1</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>4</td>
<td>12</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

An Example (continued)
\[
\begin{array}{ccccccc}
4 & 3 & 6 & 4 & 7 & 4 \\
8 & 1 & 10 & 5 & 11 & 8 \\
6 & 2 & 7 & 8 & 8 & 7 \\
9 & 4 & 12 & 9 & 12 & 9 \\
7 & 5 & 9 & 6 & 10 & 7 \\
\end{array}
\]
\[
\begin{array}{cccc}
0 & -1 & 2 & 0 & 3 \\
0 & -7 & 2 & -3 & 3 \\
-1 & -5 & 0 & 1 & 1 \\
0 & -5 & 3 & 0 & 3 \\
0 & -2 & 2 & -1 & 3 \\
\end{array}
\]

row medians
matrix after removing row medians
An Example (continued)

0   -1   2   0   3   0   4   0   0   0
0   -7   2   -3   3   0   -2   0   -3   0
-1   -5   0   1   1   -1   0   -2   1   -2
0   -5   3   0   3   0   0   1   0   0
0   -2   2   -1   3   0   3   0   -1   0
0   -5   2   0   3   

column medians

matrix after subtracting column medians

An Example (continued)

0   4   0   0   0   0   0   -2   0   -3   0   0
-1   0   -2   1   -2   -1   0   1   0   0   0   0
0   3   0   -1   0   0   

row medians

matrix after removing row medians

An Example (continued)

0   4   0   0   0   0   0   -2   0   -3   0   0
0   1   -1   2   -1   0   1   -1   2   -1   0   0
0   0   1   0   0   0   -1   1   0   0   

All row medians and column medians are 0. Thus the median polish procedure has converged. The above is the residual matrix that we will subtract from the original matrix to obtain the fitted values.

An Example (continued)

0    4    0    0    0    0
0   -2    0   -3    0   -3
0   -1   -2    1   -2   -1
0    0   -1    2   -1   0
0    3   -1    0   -1   0

column medians

matrix after subtracting column medians

An Example (continued)

4    3    6    4    7   0    3    0    0    0
8    1   10    5   11   0  -3    0   -3   0
6    2    7    8    8   0    0   -1   2   -1
0    1   -1   2    1   0    0    1    0    0
0   -1   0   -1   0   0    2   -1   0   -1
0    1    0    0    0   

matrix of fitted values

row means

R Commands for Obtaining RMA Expression Measures from Affymetrix .CEL Files

# Load the Bioconductor package affy.
library(affy)

# Set the working directory to the directory containing all the .CEL files.
setwd("C:/z/Courses/Smicroarray/AffyCel")

# Read the .CEL file data.
Data=ReadAffy()

# Compute the RMA measures of expression.
expr=rma(Data)

# Write the data to a tab-delimited text file.
write.exprs(expr, file="mydata.txt")